Statistical Methods for Improving Biosurveillance System Performance

Associate Professor Ron Fricker
Naval Postgraduate School

Army Conference on Applied Statistics
October 17, 2007

The New Status Quo?

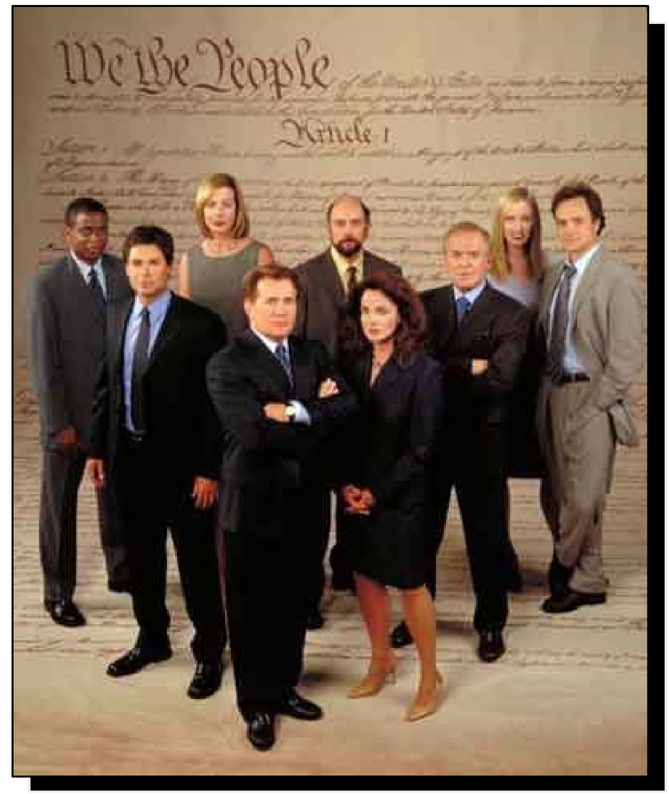
From "Emerging Health Threats and Health Information Systems: Getting Public Health and Clinical Medicine to Real Time Response"
John W. Loonsk, M.D., Associate Director for Informatics, Centers for Disease Control and Prevention

Bioterrorism Threat Recognized Pre-9/11 – Even in Popular Culture…

"That's how it's gonna be, a little test tube with a-a rubber cap that's deteriorating... A guy steps out of Times Square Station. Pshht... Smashes it on the sidewalk... There is a world war right there."

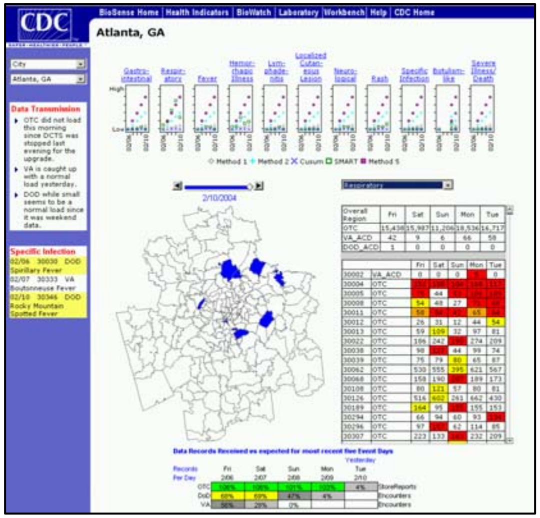
"Josh"

West Wing, 1999



Syndromic Surveillance as a Form of Biosurveillance

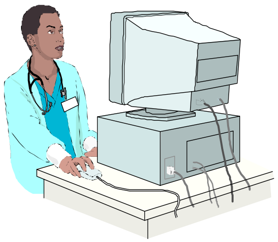
- "…surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." [1]
- On-going discussion in public health community about use of syndromic surveillance for "early event detection" vs. "situational awareness"



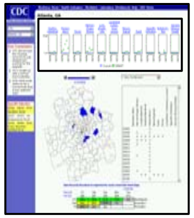
[1] CDC (www.cdc.gov/epo/dphsi/syndromic.htm, accessed 5/29/07)

Idea of Syndromic Surveillance: Leverage Secondary Health Data

- Ideal is automatic or near real-time data analysis
- Use data, methods to allow for identification of subtle trends not visible to individual MD's
- Provide indicators to trigger detection, investigation, quantification, localization, and outbreak management



Clinical Data and Lab Results



Syndromic Surveillance System

Other Early Detection Data

Derived from "Emerging Health Threats and Health Information Systems: Getting Public Health and Clinical Medicine to Real Time Response"
John W. Loonsk, M.D., Associate Director for Informatics, Centers for Disease Control and Prevention

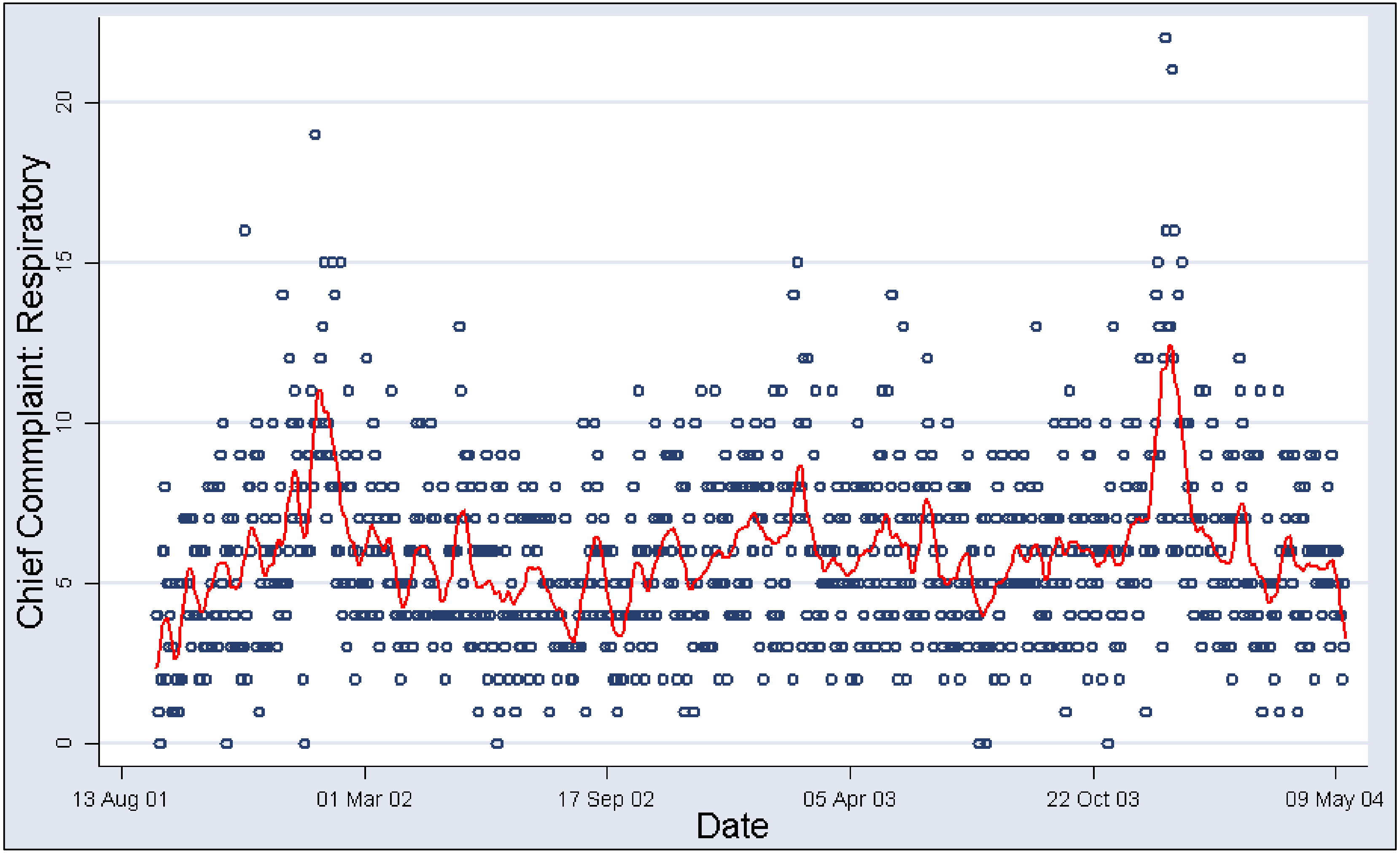
Statistical Process Control (SPC) for Syndromic Surveillance

- In manufacturing setting, SPC used to monitor production and test for a change level of quality
  - Sequential hypothesis test for distributional parameter(s) of quality characteristic (often the mean)
- In syndromic surveillance, goal is to monitor whether a pathogen has been released
  - Test whether distribution of leading indicators has shifted in some meaningful (i.e., worrisome) way
    - Focus needs to be on nonspecific—but relevant—symptoms
    - If symptoms are obvious and specific, then observation by clinician is likely sufficient

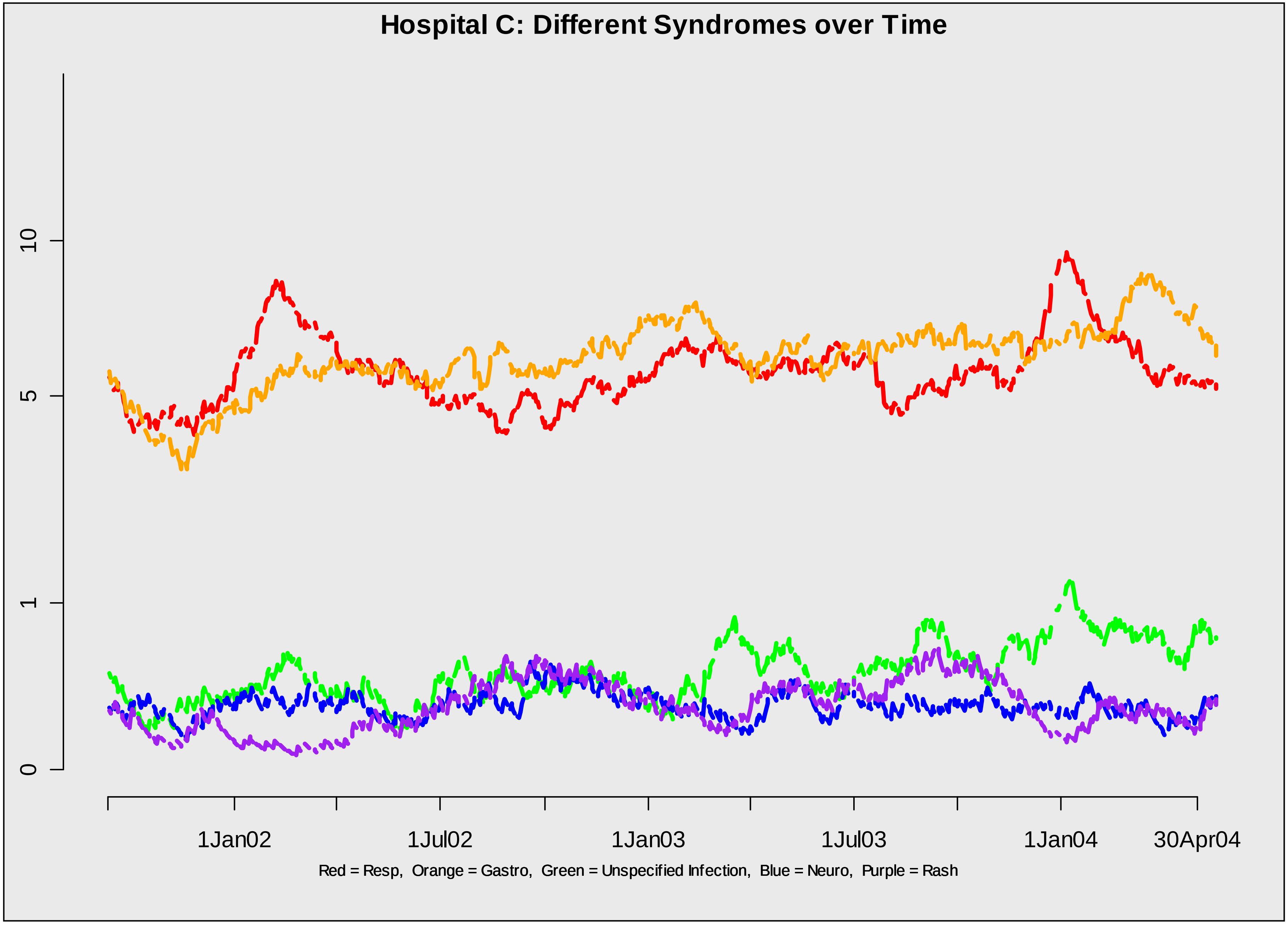
Challenges in Developing Statistical Methods for Syndromic Surveillance

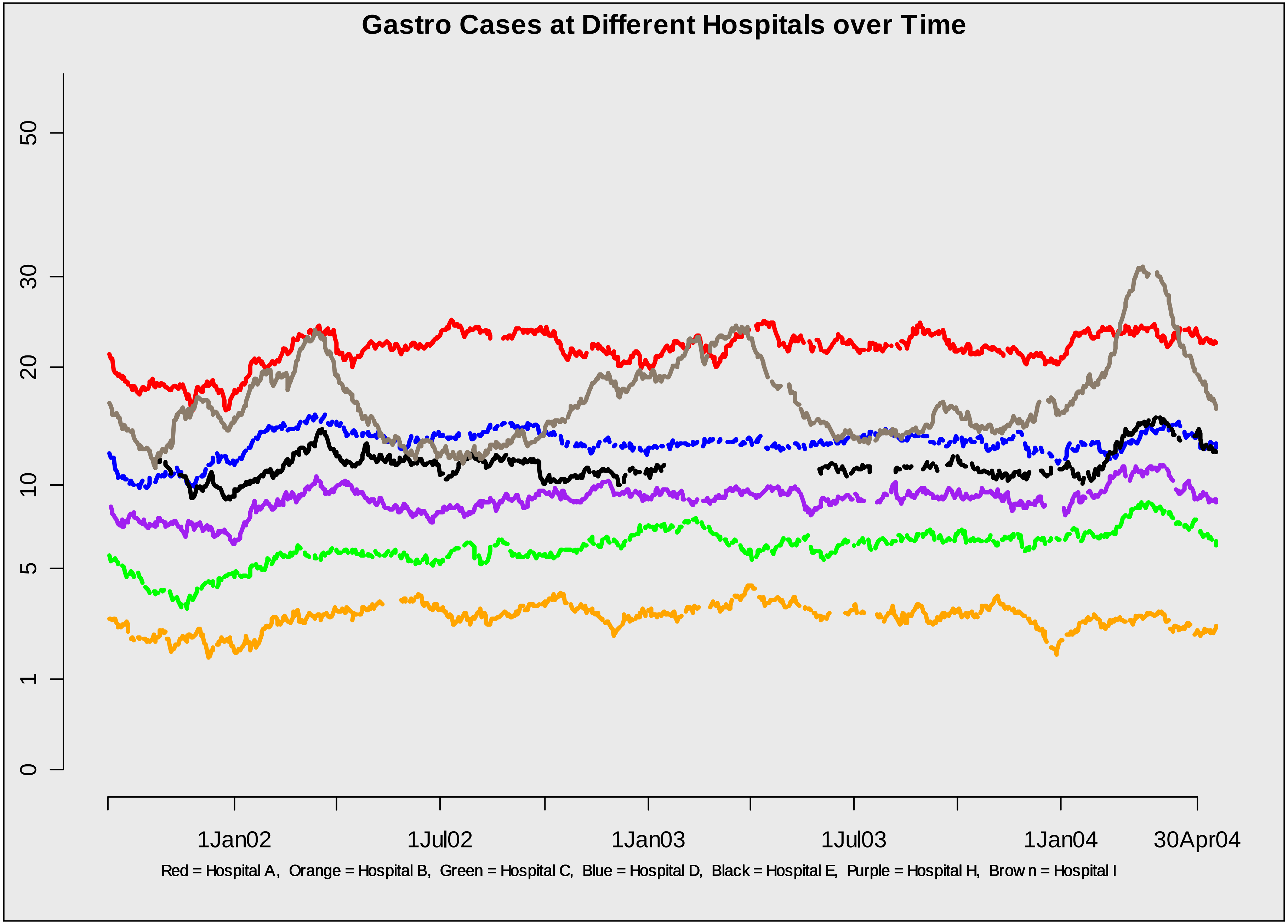
- Nonstationary data
  - No control over “in-control” distribution
- Systematic effects
  - Seasonal, day-of-the-week and other effects in data
- Transient “out-of-control” conditions
  - Outbreaks/attacks begin, peak, and subside
- Vague alternative hypotheses
  - Detect only bioterrorism or natural diseases too?
  - Which diseases and/or outbreak manifestations?

Challenges in Developing Statistical Methods for Syndromic Surveillance



Respiratory Data From "Hospital C"


Hospital C: Different Syndromes over Time
(Smoothed) Numbers of Cases
0
1
5
10
1Jan02
1Jul02
1Jan03
1Jul03
1Jan04
30Apr04
Red = Resp, Orange = Gastro, Green = Unspecified Infection, Blue = Neuro, Purple = Rash


Gastro Cases at Different Hospitals over Time
(Smoothed) Numbers of Cases
0
1
5
10
20
30
50
1Jan02
1Jul02
1Jan03
1Jul03
1Jan04
30Apr04
Red = Hospital A, Orange = Hospital B, Green = Hospital C, Blue = Hospital D, Black = Hospital E, Purple = Hospital H, Brow n = Hospital I

Improving Statistical Methods for Syndromic Surveillance

- Focus areas:
  - Improve on / adapt existing temporal methods
    - Those currently in use in biosurveillance systems
    - Statistical process control methods
  - Create and assess methods for optimizing biosurveillance systems
  - Develop / improve spatio-temporal methods

Focus Area #1: Improving on Existing Temporal Methods [1]

- Early Aberration Reporting System (EARS)

- Designed to be a “drop-in” surveillance system
  - Often little historical information available
- Available on the web[2], so increasingly being used as standard health surveillance system

[1] Based on joint work with B. Hegler and D. Dunfee
[2] www.bt.cdc.gov/surveillance/ears

EARS' Methods: C1, C2, and C3

$$C_1(t) = \frac{Y(t) - \overline{Y}_1(t)}{s_1(t)}$$

- Sample statistics calculated from previous 7 days' data
- Stop when statistic > 3

$$C_2(t) = \frac{Y(t) - \overline{Y}_3(t)}{s_3(t)}$$

- Sample statistics calculated from 7 days' of data prior to 2 day lag
- Stop when statistic > 3

$$C_3(t) = \sum_{i=t}^{t-2} \max\left[0, C_2(i) - 1\right]$$

- Stop when statistic > 2

Alternative: CUSUM on Residuals from "Adaptive Regression"

- Adaptive regression: regress a sliding baseline of observations on time relative to current observation
  - I.e. regress $Y(t-1),\ldots,Y(t-n)$ on $n,\ldots,1$
- Calculate standardized residuals from one day ahead forecast, $X(t) = R(t)/\sigma_Y$, where

$$R(t) = Y(t) - \left[\hat{\beta}_0 + \hat{\beta}_1 \times (n+1) + \hat{\beta}_j \right]$$

- CUSUM:

$$S(t) = \max\left[0, S(t-1) + X(t) - k\right]$$

with

$$k = \frac{1}{2}\sqrt{\frac{(n+2)(n+1)}{n(n-1)}}$$

Comparison Methodology

- Generate synthetic data:

$$Y(t) = \max\left(0, \left\lceil c + s(t) + d(t) + Z(t) + o(t) \right\rceil\right)$$

- Scenarios:

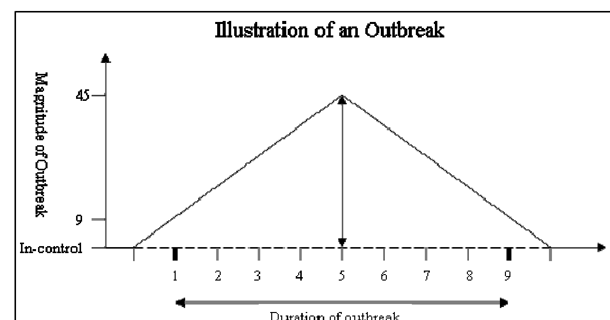
	None	*Small*	*Large*
A	0	20	80
σ	n/a	10	30

Large count: c=90

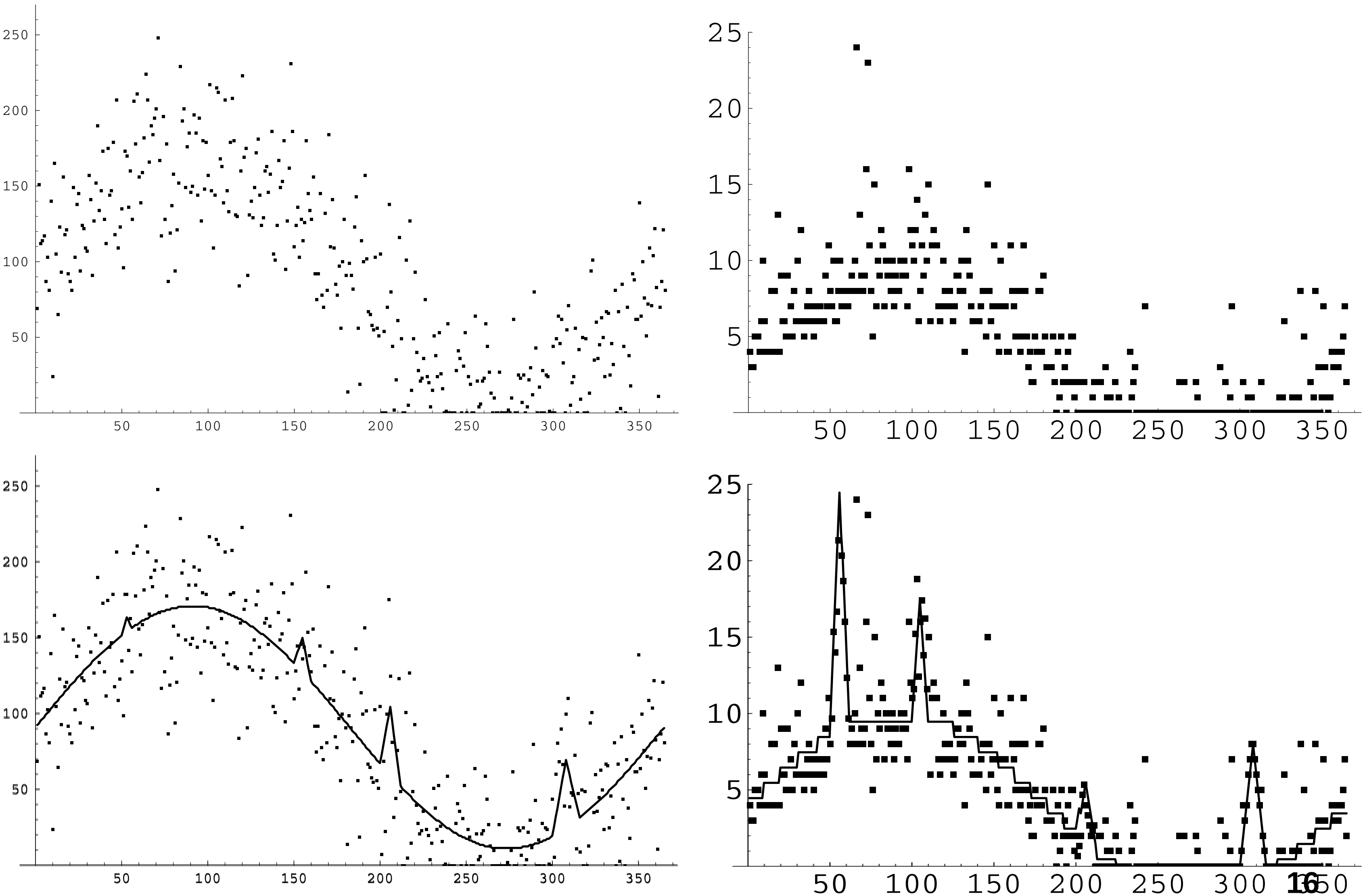
	None	*Small*	*Large*
A	0	2	6
μ, σ	n/a	1.0, 0.5	1.0, 0.7

Small count: c=0

- Outbreaks
  - Linear increase & decrease
  - Characterized by duration and magnitude



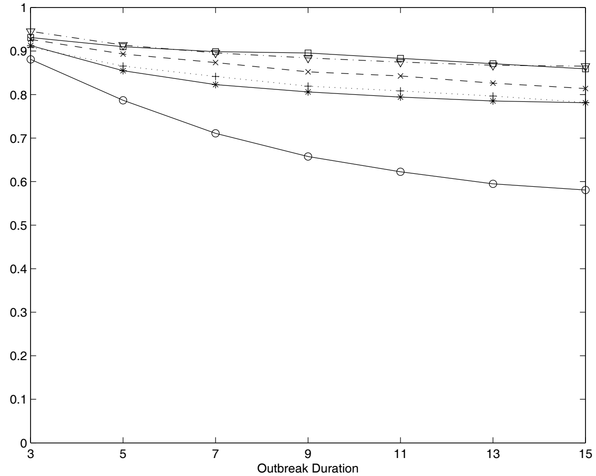
Synthetic Data: Outbreaks?

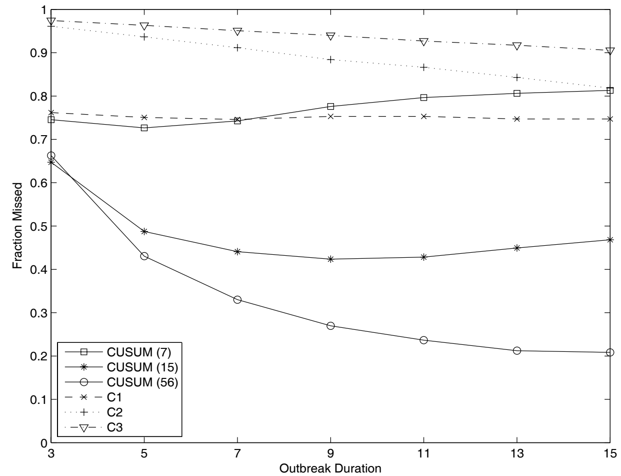


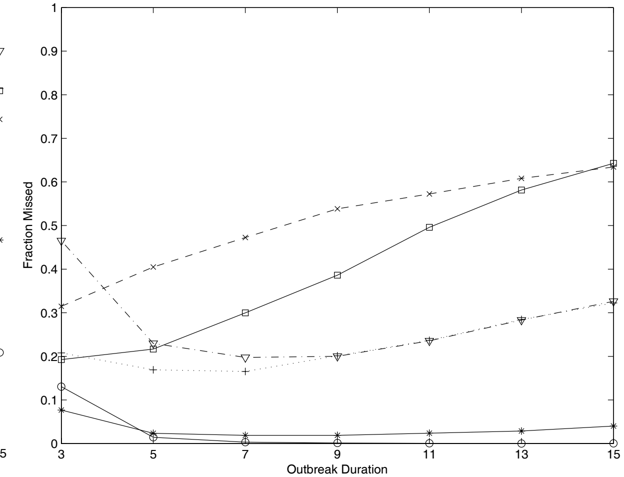
Some Large Count Results

Small magnitude | Medium magnitude | Large magnitude

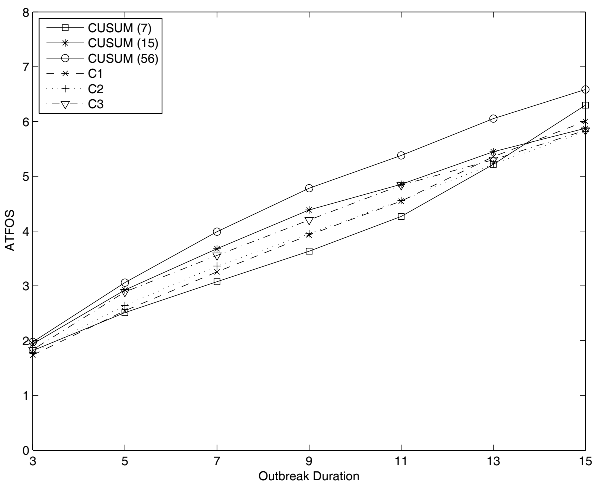
Fraction Missed

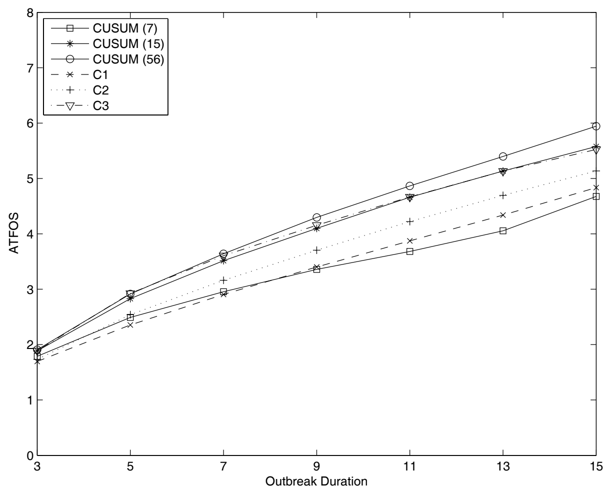


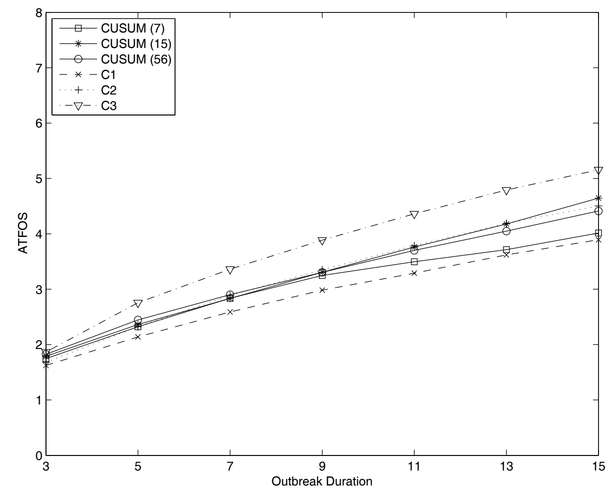




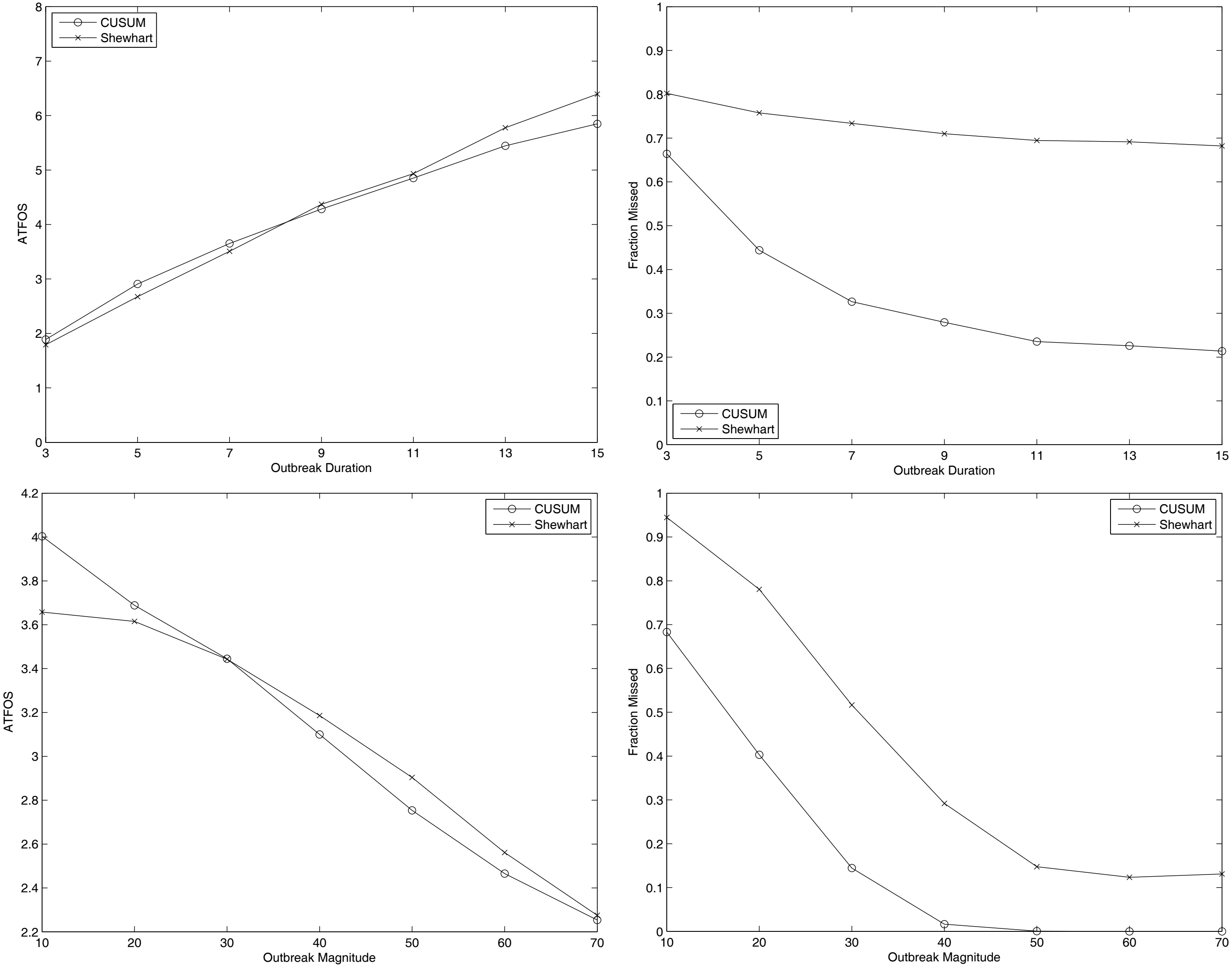
Avg Time to Signal







Shewhart Methods Not Suited for this Problem?



Observations

- CUSUMs based on adaptive regression with longer baselines performed best
- CUSUMs outperformed EARS' methods
  - Seemingly due to Shewhart design *and* additional data used in adaptive regression
- Suggests "drop in" strategy of starting with CUSUM with 7-day baseline
  - As time progresses, increase baseline until long enough to allow it to slide

Focus Area #2: Optimizing Biosurveillance Systems [1]

- False alarms a serious problem

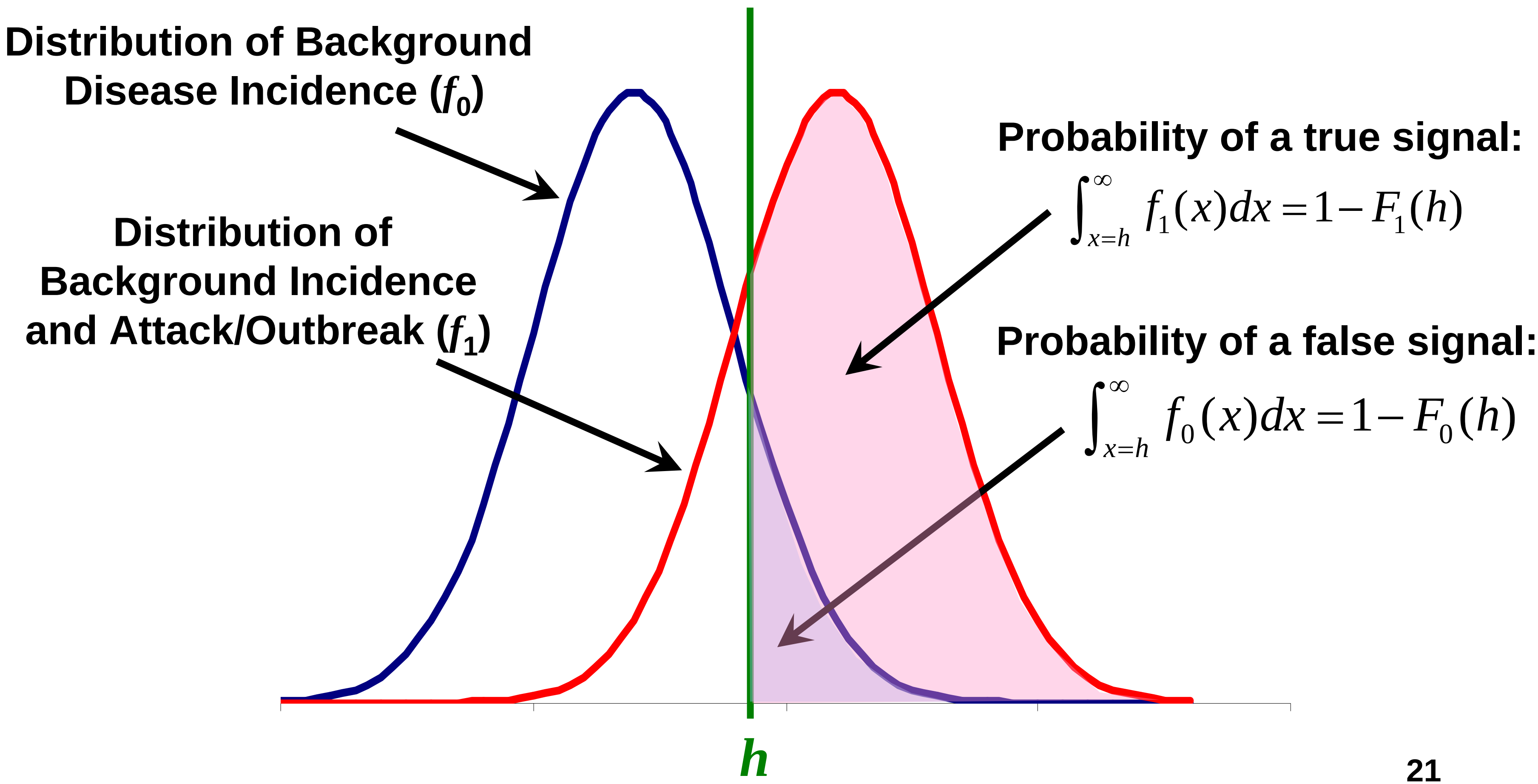
  "…most health monitors, who are the users of such systems, learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!" [2]

- Issue: How to optimize the <u>system</u>?
  - I.e., How to choose thresholds to maximize the probability of detecting an attack within a manageable false alarm rate

[1] Based on joint work with D. Banschbach
[2] https://wiki.cirg.washington.edu/pub/bin/view/Isds/SurveillanceSystemsInPractice

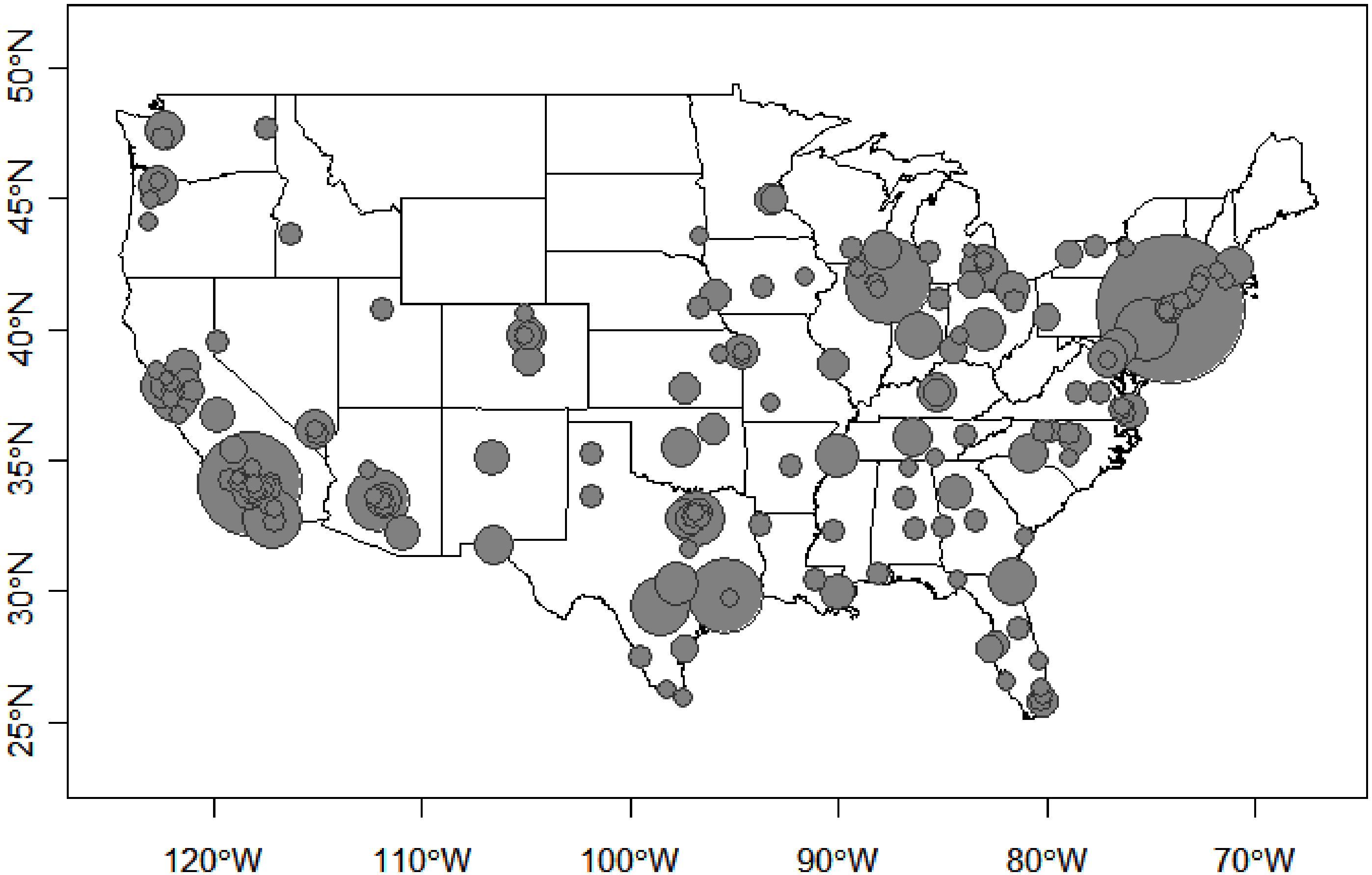
Threshold Detection Methodology

- Signal if observation exceeds threshold h



Consider (Hypothetical) System to Monitor 200 Largest Cities in US

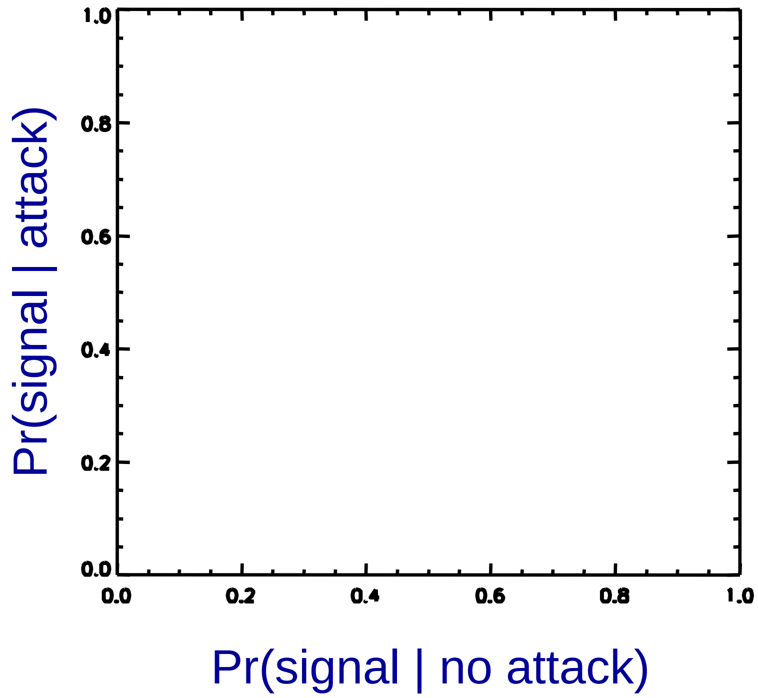
- Assume probability of attack is proportional to the population in a city: $p_i = m_i / \sum_i m_i$

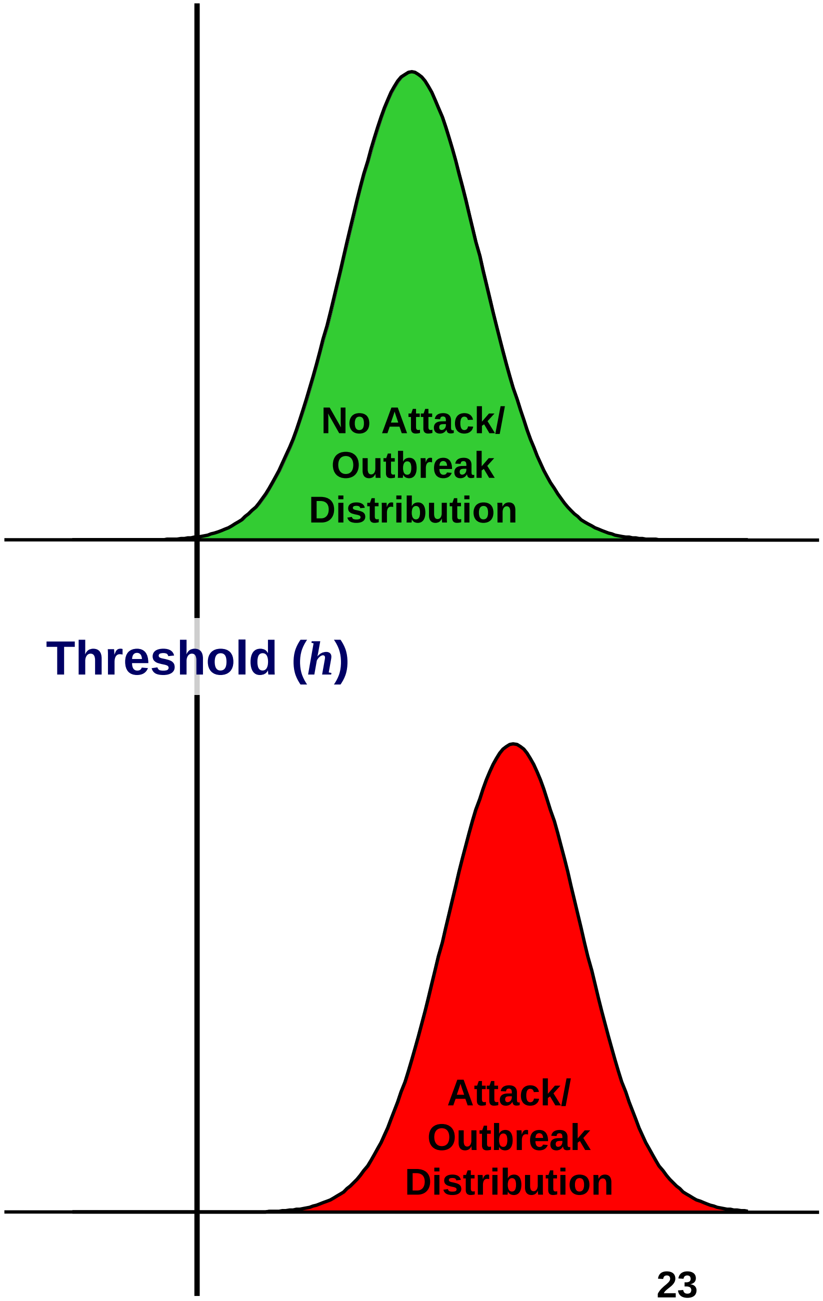


It's All About Choosing Thresholds

- For each city, choice of h is compromise between probability of true and false signals

ROC Curve





Mathematical Formulation of the Problem

- It's simple to write out:

$$\Pr(\text{detection}) = \sum_i \Pr(\text{signal}|\text{attack}) \Pr(\text{attack})$$

$$E(\#\text{ false signals}) = \sum_i \Pr(\text{signal}|\text{no attack})$$

- Express it as an optimization problem:

$$\max_{\vec{h}} \sum_i \left[1 - F_1(h_i)\right] p_i$$

Choose thresholds to maximize the probability of detecting an attack

$$\text{s.t.} \sum_i \left[1 - F_0(h_i)\right] \leq K$$

Subject to a constraint on the expected number of false alarms

Visualizing an Objective Function and Feasible Region

- Consider a simple two-sensor problem:
  - F_0=N(0,1) & F_1=N(1,1)
  - Equal probability of attack at either sensor
  - Constraint: $\alpha_1+\alpha_2\leq 0.2$
- In general, global optimal solution exists if, over the feasible region:
  - F_0 and F_1 concave
  - F_1 strictly increasing

Figure 1 Plot of an objective function for $n=2$ with $F_0=N(0,1)$, $F_1=N(1,1)$ and $\mathbf{p}=\{1/2,1/2\}$.

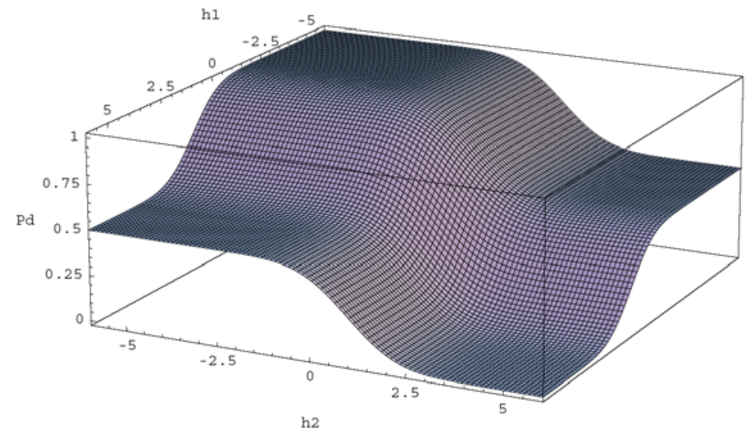
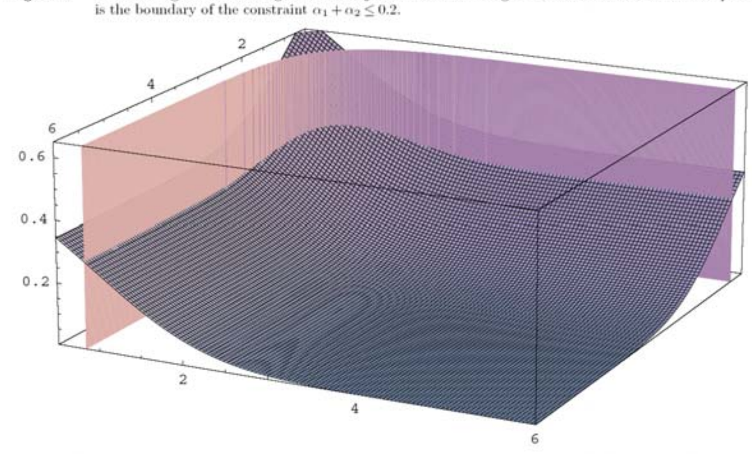
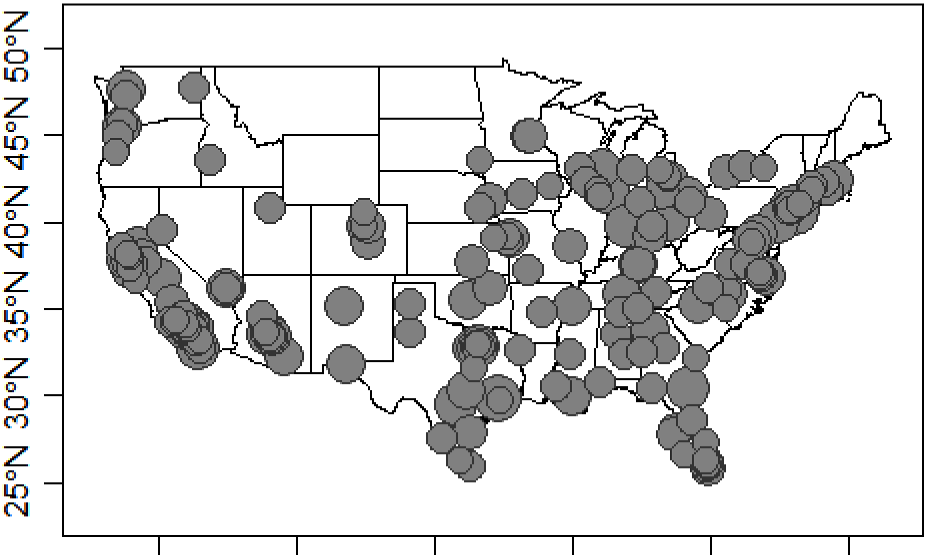


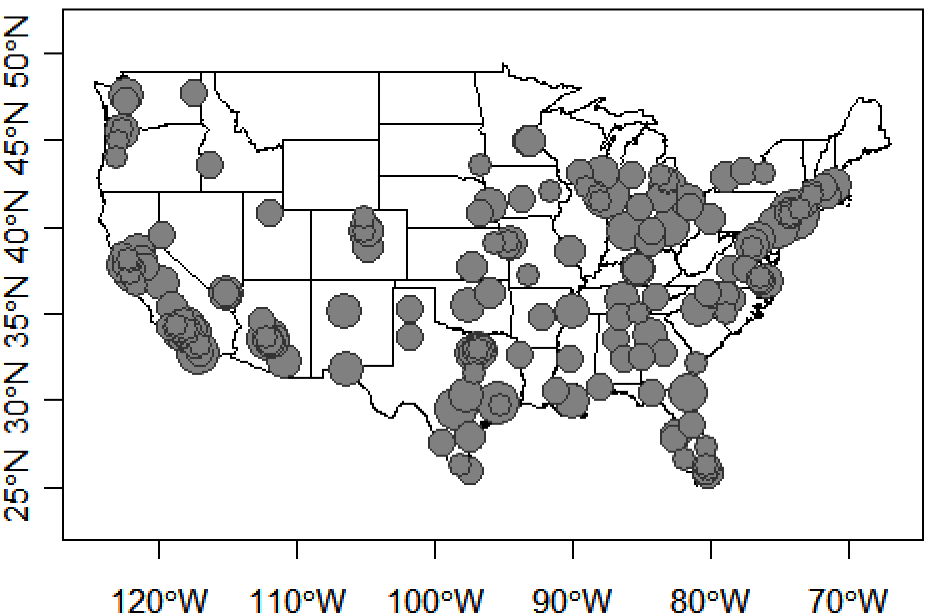
Figure 2 Plot showing the feasible region of the objective function in Figure 1, where the vertical curved plane is the boundary of the constraint $\alpha_1+\alpha_2\leq 0.2$.





Probability of detection = 0.5
Expected number of false alarms = 4.55

Probability of detection = 0.6
Expected number of false alarms = 4.55

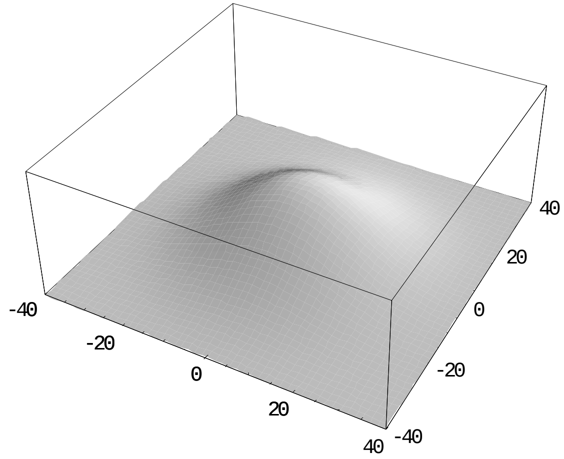


Probability of detection = 0.5
Expected number of false alarms = 4.55

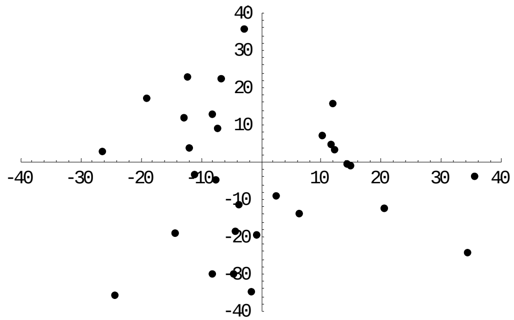
Probability of detection = 0.5
Expected number of false alarms = 2.25

Focus Area #3: Spatio-temporal Methods (Detection and Situational Awareness) [1]

- ER patients come from surrounding area
  - On average, 30 per day
    - More likely from closer distances
  - Outbreak occurs at (20,20)
    - Number of patients increase linearly by day after outbreak



(Unobservable) distribution of ER patients' home addresses

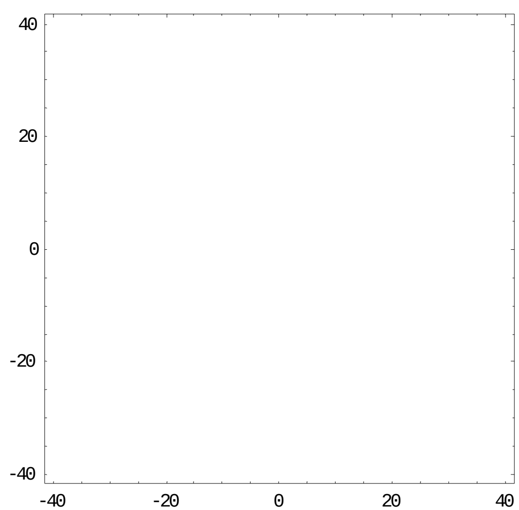


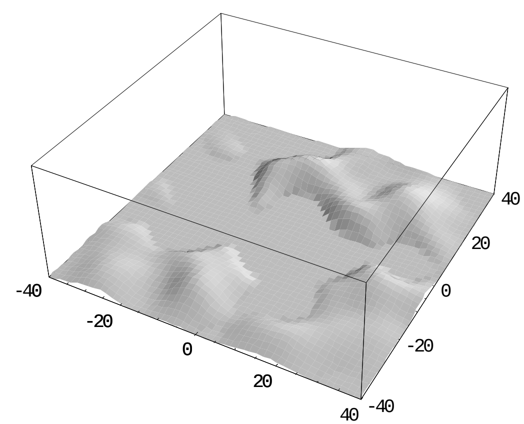
Observed distribution of ER patients' home addresses

[1] Based on joint work with J. Chang

Use Difference in Kernel Density Estimates to Assess Outbreak Location

- Construct kernel density estimate of “normal” disease incidence using N historical observations
- Compare to density estimate of most recent w observations





But how to know when to signal?

Idea: Conduct Sequential Hypothesis Test of Distribution of Density Heights

- At each time *n*, use historical data and recent data to construct density estimate, $\hat{f}_n$
  - Under no-outbreak condition, density heights of historical and recent observations should be uniformly distributed
- Compute empirical distributions of the two:

$$\hat{J}_n(z) = \frac{1}{w+1} \sum_{i=n-w}^{n} I\left\{ \hat{f}_n(X_i) \le z \right\}, \qquad \hat{H}_n(z) = \frac{1}{N} \sum_{i=n-w-N-1}^{n-w-1} I\left\{ \hat{f}_n(X_i) \le z \right\}$$

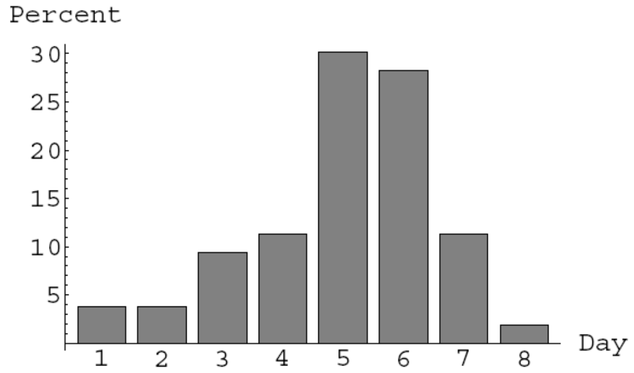
- Use Kolmogorov-Smirnov test to assess when differences are significant:

$$S_n = \max_z \left| \hat{J}_n(z) - \hat{H}_n(z) \right|$$

  - Signal at time $t = \min\left\{ n : S_n > h \right\}$

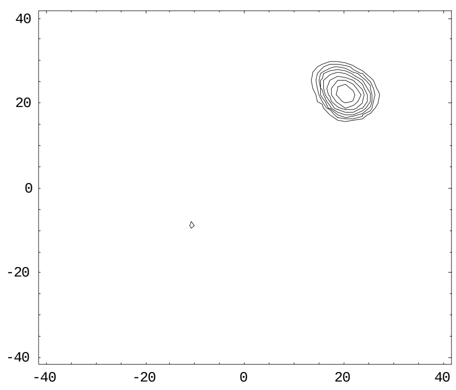
Illustrative Results

- Assess performance by simulating outbreak multiple times, record when RTR signals
  - Signaled middle of day 5 on average
  - By end of 5th day, 15 outbreak and 150 non-outbreak observations

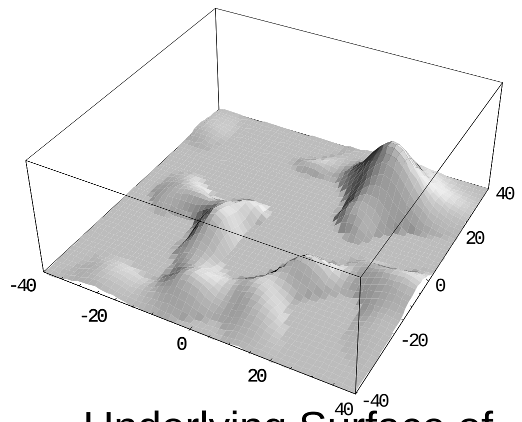


Distribution of Signal Day

  - From previous example:



Estimate of Outbreak Location on Day 5



Underlying Surface of Density Height Differences

In Summary: Lots of Interesting Research Opportunities

- Control chart experts have a lot to offer
  - Public health community re-inventing SPC?
  - Guidance for how to compare methods
  - Determining metrics for biosurveillance methods
- Syndromic surveillance offers challenges
  - Adapting/extending existing methods
  - Developing and assessing new
    - Univariate and multivariate temporal methods
    - Spatio-temporal methods
  - Strategies for optimizing biosurveillance systems

Selected References

Background Information:

- Woodall, W.H., The Use of Control Charts in Health-Care and Public-Health Surveillance, *Journal of Quality Technology*, **38**, pp. 1-16, 2006.
- Fricker, R.D., Jr., Syndromic Surveillance, *Encyclopedia of Quantitative Risk Assessment* (to appear).
- Fricker, R.D., Jr., and H. Rolka, Protecting Against Biological Terrorism: Statistical Issues in Electronic Biosurveillance, *Chance*, **91**, pp. 4-13, 2006
- Shmueli, G., Statistical Challenges in Modern Biosurveillance, in submission to *Technometrics*.

Selected Research:

- Fricker, R.D., Jr., Hegler, B.L., and D.A Dunfee, Assessing the Performance of the Early Aberration Reporting System (EARS) Syndromic Surveillance Algorithms, in submission to *Statistics in Medicine*.
- Fricker, R.D., Jr., and D. Banschbach, Optimizing a System of Threshold Detection Sensors, in draft.
- Fricker, R.D., Jr., and J.T. Chang, A Spatio-temporal Method for Real-time Biosurveillance, in submission to *Quality Engineering*, 2007.
- Fricker, R.D., Jr., Knitt, M.C., and C.X. Hu, Comparing Directionally Sensitive MCUSUM and MEWMA Procedures with Application to Biosurveillance, in submission to *Quality Engineering*.
- Joner, M.D., Jr., Woodall, W.H., Reynolds, M.R., Jr., and R.D. Fricker, Jr., A One-Sided MEWMA Chart for Health Surveillance, in submission to *Quality and Reliability Engineering International*.
- Fricker, R.D., Jr., Directionally Sensitive Multivariate Statistical Process Control Methods with Application to Syndromic Surveillance, *Advances in Disease Surveillance,* **3**:1.